EXPRESS MAIL NO.
EV064992167US

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 March 2001 (29.03.2001)

PCT

(10) International Publication Number
WO 01/22280 A2

(51) International Patent Classification[7]: G06F 17/30

(21) International Application Number: PCT/IL00/00582

(22) International Filing Date:
19 September 2000 (19.09.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/399,618 20 September 1999 (20.09.1999) US

(71) Applicant *(for all designated States except US)*: CLEARFOREST LTD. [IL/IL]; P.O. Box 786, 60256 Or Yehuda (IL).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: FELDMAN, Ronen [IL/IL]; Halevi Street 12, 49559 Petach Tikva (IL). AUMANN, Yehonatan [IL/IL]; Ben Maimon Avenue 39, 92262 Jerusalem (IL). BEN-YEHUDA, Yaron [IL/IL]; Had-Nes Street 36, 52331 Ramat Gan (IL). LANDAU, David [IL/IL]; Ezra Street 10, 76200 Rehovot (IL).

(74) Agents: COLB, Sanford, T. et al.; Sanford T. Colb & Co., P.O. Box 2273, 76122 Rehovot (IL).

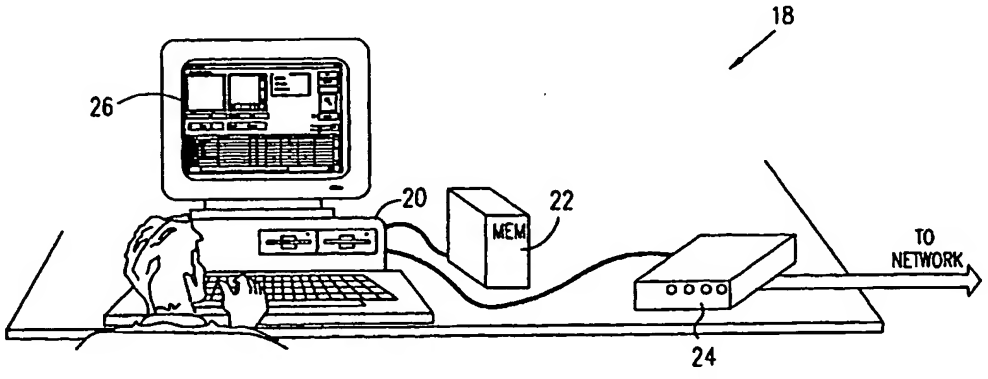
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *Without international search report and to be republished upon receipt of that report.*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: DETERMINING TRENDS USING TEXT MINING



(57) Abstract: A method for visualizing variations in a corpus of information. The corpus includes a plurality of information entries, which are divided into a plurality of sub-groups according to a differentiating parameter of the entries. For each of the entries, characteristics of information contained therein are extracted and pairs of different characteristics that appear together in at least one of the entries are found. An occurrence value is determined for each of the pairs of characteristics in each sub-group in which both of the characteristics appear. The occurrence values of at least some of the pairs of characteristics for at least two of the sub-groups are compared, and an indication of the comparative occurrence values of the pairs is provided.

WO 01/22280 A2

DETERMINING TRENDS USING TEXT MINING

FIELD OF THE INVENTION

The present invention relates generally to knowledge discovery in collections of data, and specifically to text mining.

BACKGROUND OF THE INVENTION

In recent years, the volume of text documents available on computers and computer networks is growing rapidly. It is virtually impossible to read all the available documents containing information of importance on a given subject. In order to find desired information, search engines have been developed which provide a user with documents which mention selected words or terms. The user may use Boolean patterns with "and," "or" and "not" terms to more distinctly define the scope of the desired documents. However, the user cannot always define precisely which are the desired documents or keyword combinations. In addition, search engines do not provide an integrated picture of the distribution and impact of given terms in an entire corpus of documents.

Text mining is used to find hidden patterns in large textual collections. Text mining tools provide a human-tangible description of the information included in the textual collection. Because the amount of information is so large, a crucial feature of text mining tools is the way the information is organized and/or displayed. To limit the amount of information that a user must digest, it is common to define a context group which defines the information of interest to the particular user. Normally, the context group includes those documents which include one or more terms from a user-defined set.

A central tool in text mining is visualization of the complex patterns that are discovered. One such visualization approach is described, for example, in an article by Feldman R., Klosgen W., and Zilberstien A., entitled "Visualization Techniques to Explore Data Mining Results for Document Collections," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* (1997), pp. 16-23, which is incorporated herein by reference. This article describes a concept relationship analysis in which a set of concepts (or terms) are searched for in a corpus of textual data formed of a plurality of documents. The concept relationship analysis searches for groups

of concepts which appear together in relatively large numbers of documents, and these concepts are displayed together.

One method of representing concept relationships is by displaying context graphs. In context graphs, the concepts (or terms) which appear together in large numbers of documents are designated by nodes. Each two nodes are connected by an edge which has a weight which is equal to the number of documents in which the terms of both nodes appear together. In order to limit the amount of data displayed, only edges which have a weight above a predetermined threshold are displayed. In some context graphs, the concepts which appear in nodes are chosen from a list of interesting terms defined by the user.

In many cases, the corpus of documents is formed of several groups of documents, for example, documents from different dates, and it is desired to apprehend concept relationships as they develop in time. An article by Lent B., Agrawal R., and Srikant R., entitled "Discovering Trends in Text Databases," in *Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining* (1997), pp. 227-230, which is incorporated herein by reference, describes a method of detecting trends in textual collections formed of documents with timestamps, which are partitioned into time groups according to a selected granularity. The textual collection is mined for a group of combinations of words (referred to as phrases) which appear in the documents of the collection. Each combination is given frequency-of-occurrence values for each time group. A user requests to view the frequencies of occurrence of those combinations for which the occurrences follow a desired pattern. However, this method does not give the user any feel for the development of trends in the textual documents as a whole.

In an article entitled "Trend graphs: Visualizing the evolution of concept relationships in large document collections," by Feldman R., Aumann Y., Zilberstien A., and Ben-Yehuda Y., in *Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining* (1998), which is incorporated herein by reference, a graphical tool is described for analyzing and visualizing dynamic changes in concept relationships over time.

SUMMARY OF THE INVENTION

It is an object of the present invention to provide methods and apparatus for displaying trends that are discovered in large collections of information.

In some aspects of the present invention, the trends relate to appearances of terms found by text mining in groups of documents.

It is another object of some aspects of the present invention to provide methods and apparatus for displaying the evolution of concept relationships in groups of documents.

It is another object of some aspects of the present invention to provide methods and apparatus for displaying differences between patterns of term appearances in different groups of documents.

It is still another object of some aspects of the present invention to provide methods and apparatus for determining major changes in patterns of term appearances in groups of documents.

In preferred embodiments of the present invention, a corpus of documents is divided into sub-groups defined by a differentiating parameter, such as the dates of the documents, or their origin. For each sub-group of documents, a separate context graph is prepared, and the relationship between the graphs is calculated.

In some preferred embodiments of the present invention, the differentiating parameter defines an order of the context graphs. The context graphs are preferably displayed sequentially, either one after another or one above the other. Each graph is preferably displayed with indications which show the differences between the present graph and the previous graph. Preferably, each edge in the graph is marked to indicate a difference between its weight in the present graph and its weight in the previous graph. Alternatively or additionally, each edge is marked to indicate the difference between its weight in the present graph and its average weight in a predetermined number of previous graphs.

Preferably, the edges are marked graphically, for example, using different colors, widths, and/or lengths to indicate the weight differences. In a preferred embodiment of the present invention, four indications are used for the following groups of edges: new edges,

edges with increased weights, edges with decreased weights, and edges with substantially stable weights.

In some preferred embodiments of the present invention, the differentiating parameter is the date of the documents. Preferably, all the documents from a single period are considered to belong to a single sub-group. The periods may be of substantially any length, e.g., from minutes to years, according to a user selection. Alternatively or additionally, the differentiating parameter comprises the origins of the documents, such as the authors, editors, countries of origin or the original languages of the documents. Further alternatively or additionally, substantially any other parameter may be used, such as the length of a document, or the average salary or number of employees of the company mentioned most frequently in a document.

In a preferred embodiment of the present invention, the context graphs are displayed such that all nodes that are common to two or more of the graphs appear in substantially the same relative locations in the graphs. Therefore, the layout of the displayed form of the context graphs is prepared after all the nodes of all the graphs are known. Alternatively, the locations of the nodes and/or the distances between the nodes are used to indicate the importance of the terms of the nodes. In such cases, animation techniques are preferably used to aid the user to follow the changes in the positions of the nodes.

In some preferred embodiments of the present invention, an animation sequence is used to display the changes between the context graphs. Alternatively or additionally, the context graphs are listed, for example, in a list box, and the user can choose which context graph should be displayed relative to which other graphs. Further alternatively or additionally, a plurality of context graphs are superimposed one over the other, and each graph is displayed using a different color.

In some preferred embodiments of the present invention, the corpus of documents includes a set of documents selected by a search engine, a clustering program, or by any other method of filtering and/or gathering of documents. Furthermore, the trend graphs produced in accordance with preferred embodiments of the present invention may be used to select groups of documents on which additional filtering and/or other processing is to be performed.

Although preferred embodiments are described herein with reference to mining and analysis of text documents, those skilled in the art will appreciate that the principles of the present invention may also be applied to visualization of trends and other variations in collections of information of other types. For example, trends occurring among the records in a large database may be analyzed and visualized in similar fashion.

There is therefore provided, in accordance with a preferred embodiment of the present invention, a method for visualizing variations in a corpus of information, including a plurality of information entries which are divided into a plurality of sub-groups according to a differentiating parameter of the entries, including:

for each of the entries, extracting characteristics of information contained therein;

finding pairs of different characteristics that appear together in at least one of the entries;

determining an occurrence value for each of the pairs of characteristics in each sub-group in which both of the characteristics appear;

comparing the occurrence values of at least some of the pairs of characteristics for at least two of the sub-groups; and

providing an indication of the comparative occurrence values of the pairs.

Preferably, the entries include text documents, and the characteristics include terms appearing in the documents.

Further preferably, determining the occurrence value includes counting the number of entries in which the pair appears.

Still further preferably, finding the pairs of characteristics includes finding pairs of characteristics which appear together in at least a predetermined number of the entries.

In a preferred embodiment, finding the pairs of characteristics includes finding pairs of characteristics which appear together in at least two of the sub-groups.

Preferably, extracting the characteristics includes automatically mining the corpus to extract characteristics therefrom.

In a preferred embodiment, the differentiating parameter defines an order, and comparing the occurrence values includes comparing the occurrence values in a first sub-group with the occurrence values in one or more previous sub-groups in the order. Preferably, comparing the occurrence values includes comparing the occurrence values in the first sub-group with the occurrence values in a closest previous sub-group.

Alternatively or additionally, comparing the occurrence values includes comparing the occurrence values in the first sub-group with an average of the occurrence values in the one or more previous sub-groups. Further alternatively or additionally, providing the indication includes displaying a symbol which indicates a measure of evolution in the occurrence value in the first sub-group relative to the occurrence values in the one or more previous sub-groups in the order.

In a preferred embodiment, providing the indication includes displaying a table or graph. Preferably, displaying the graph includes displaying a graph in which each term is represented by a node, the pairs of characteristics that are found are represented by edges, and substantially each edge is associated with the indication of the comparative appearance of the respective pair. Typically, displaying the graph includes displaying with substantially each edge a weight of the edge, which equals the occurrence value of the respective pair in a first sub-group. Alternatively or additionally, displaying the graph includes displaying the graph such that the lengths of the edges represent the occurrence value of the respective pair in a first sub-group.

In a preferred embodiment, displaying the graph includes displaying for each two sub-groups a graph which compares the occurrence values in the two sub-groups. Preferably, displaying the graph for each two sub-groups includes displaying the graphs such that nodes which represent the same term are displayed in substantially the same relative location. Further preferably, the graphs of each two sub-groups are displayed as an animation sequence.

Preferably, displaying the graph includes displaying a plurality of superimposed graphs, each of which represents the appearances of the pairs in a different sub-group. Further preferably, displaying the plurality of superimposed graphs includes displaying each of the graphs in a different color.

In a preferred embodiment, providing the indication of the comparative values of the pairs includes providing an indication wherein which pairs having a characteristic in common are grouped together.

There is also provided, in accordance with a preferred embodiment of the present invention, apparatus for visualizing variations in a corpus of information including a plurality of information entries which are divided into a plurality of sub-groups according to a differentiating parameter of the entries, including:

a processor which finds pairs of characteristics which appear together in at least one of the documents, determines an occurrence value for each of the pairs of characteristics in each sub-group in which both of the characteristics appear, and compares the occurrence values of at least some of the pairs of characteristics for at least two of the sub-groups; and

a display which displays an indication of the comparative occurrence values of the pairs.

In a preferred embodiment, the processor finds characteristics selected from a group of automatically determined characteristics.

There is further provided, in accordance with a preferred embodiment of the present invention, a method for selecting a range of values of a variable, including:

providing a graphic user interface on a display, including a slide-piece that has an initial dimension and is translatable along an axis representing the variable such that each position of the slide-piece along the axis corresponds to a given value of the variable;

positioning the slide-piece at a first position on the axis, so as to indicate a first value of the variable; and

changing the dimension of the slide-piece so as to indicate a second value of the variable, whereby the first and second values of the variable define the selected range.

Preferably, changing the dimension of the slide-piece includes changing a length of the slide-piece along the axis. Further preferably, the first and second values of the variable include the extrema of the range.

There is still further provided, in accordance with a preferred embodiment of the present invention, a computer program product for visualizing variations in a corpus of information, including a plurality of information entries which are divided into a plurality of sub-groups according to a differentiating parameter of the entries, the documents including text, the program having computer-readable program instructions embodied therein, which instructions cause a computer to:

for each of the entries, extract characteristics of information contained therein;

find pairs of different characteristics that appear together in at least one of the entries;

determine an occurrence value for each of the pairs of characteristics in each sub-group in which both of the characteristics appear;

compare the occurrence values of at least some of the pairs of characteristics for at least two of the sub-groups; and

provide an indication of the comparative occurrence values of the pairs.

There is also provided, in accordance with a preferred embodiment of the present invention, a computer program product for selecting a range of values of a variable, the program having computer-readable program instructions embodied therein, which instructions cause a computer to:

provide a graphic user interface on a display, including a slide-piece that has an initial dimension and is translatable along an axis representing the variable such that each position of the slide-piece along the axis corresponds to a given value of the variable;

position the slide-piece at a first position on the axis, so as to indicate a first value of the variable; and

change the dimension of the slide-piece so as to indicate a second value of the variable, whereby the first and second values of the variable define the selected range.

There is additionally provided, in accordance with a preferred embodiment of the present invention, a method for extracting data from a corpus of information, including a plurality of information entries, each entry being assigned to one or more sub-groups according to a differentiating parameter of the entries, including:

for a first one of the entries in a first one of the sub-groups, extracting a characteristic of information contained therein;

for a second one of the entries in a second one of the sub-groups, extracting the same characteristic of information;

automatically determining respective first and second occurrence values corresponding to the characteristic in the first and second sub-groups; and

providing an indication of the occurrence values.

Preferably, providing the indication includes providing a visual indication of the occurrence values. Further preferably, the differentiating parameter includes a sequence, most preferably a time sequence.

There is still additionally provided, in accordance with a preferred embodiment of the present invention, apparatus for extracting data from a corpus of information including a plurality of information entries, each entry being assigned to one or more sub-groups according to a differentiating parameter of the entries, including:

a processor, which (a) for a first one of the entries in a first one of the sub-groups, extracts a characteristic of information contained therein, (b) for a second one of the entries in a second one of the sub-groups, extracts the same characteristic of information, and (c) automatically determines respective first and second occurrence values corresponding to the characteristic in the first and second sub-groups; and

a display, which provides an indication of the occurrence values.

There is yet additionally provided, in accordance with a preferred embodiment of the present invention, a computer program product for extracting data from a corpus of information, including a plurality of information entries, each entry being assigned to one or more sub-groups according to a differentiating parameter of the entries, the program having computer-readable program instructions embodied therein, which instructions, when read by a computer, cause the computer to:

for a first one of the entries in a first one of the sub-groups, extract a characteristic of information contained therein;

for a second one of the entries in a second one of the sub-groups, extract the same characteristic of information;

automatically determine respective first and second occurrence values corresponding to the characteristic in the first and second sub-groups; and

provide an indication of the occurrence values.

The present invention will be more fully understood from the following detailed description of the preferred embodiments thereof, taken together with the drawings in which:

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a schematic illustration of a system for text mining, in accordance with a preferred embodiment of the present invention;

Fig. 2 is a flow chart illustrating preparation of a trend graph from a corpus of documents, in accordance with a preferred embodiment of the present invention;

Fig. 3 is a schematic view of a text mining input window display, in accordance with a preferred embodiment of the present invention;

Fig. 4A is a schematic view of a trend graph, in accordance with a preferred embodiment of the present invention;

Fig. 4B is a schematic view of a trend graph representing a period following the period represented by the graph of Fig. 4A, in accordance with a preferred embodiment of the present invention;

Fig. 5 is a schematic view of a comparison graph, in accordance with a preferred embodiment of the present invention; and

Fig. 6 is a schematic view of a graphic interface, in accordance with a preferred embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

Fig. 1 is a schematic illustration of a system 18 for text mining and visualization, in accordance with a preferred embodiment of the present invention. System 18 preferably comprises a memory 22, which stores a corpus of documents from which information is mined. Alternatively or additionally, system 18 comprises a modem 24 or other network connection, through which access is established to collections of documents, which include some or all of the documents in the corpus. System 18 preferably further comprises a computer 20, which mines information from the documents, and a display 26, on which the mined information is displayed.

Fig. 2 is a flow chart illustrating the actions of computer 20 in preparing trend graphs from the corpus of documents, in accordance with a preferred embodiment of the present invention. Preferably, the documents in the corpus are dated and/or time-stamped, and the trend graph represents changes in the corpus as a function of time. Alternatively or additionally, each document is associated with a different ordering parameter value, not necessarily time-related. For example, the corpus of documents may include articles drawn from *The Wall Street Journal* about high-tech companies, and the ordering parameter may be the average employee salary or the number of employees of the company mentioned most frequently in an article. In this example, a database containing information about employees of high-tech companies would preferably be accessible, either locally or remotely, to computer 20. Alternatively or additionally, a more complex ordering parameter, such as (average employee salary) * (percentage of employees who use a PC) * (percentage of employees who have a college degree), may be used to aid a user in analyzing a very large collection of news articles.

Preferably, computer 20 analyzes each document and prepares for each document a record which represents the document. The record preferabiy comprises a set of terms

which appear in the document, most preferably together with the numbers of occurrences of the terms and/or a parameter which represents the importance of the terms. The records are preferably prepared in accordance with the method described in the article by Feldman, Klosgen, and Zilberstien, which is referenced in the Background of the Invention section of the present patent application. Alternatively or additionally, term extraction methods, term processing methods, and/or graphical display methods described in co-pending US patent application 09/323,491, "Term-Level Text Mining with Taxonomies," filed June 1, 1999, which is assigned to the assignee of the present patent application and is incorporated herein by reference, are used in implementing some embodiments of the present invention.

The records are preferably stored in memory 22 for future text mining, and computer 20 preferably does not need to access the documents again in order to perform additional text mining sessions.

Reference is also made to Fig. 3, which is a schematic view of a text mining input window 38 on display 26, in accordance with a preferred embodiment of the present invention. In defining a text mining session, the user preferably defines a context group on which the session is performed. Preferably, the context group comprises those documents in which one or more selected terms appear in the accumulative or in the alternative, according to user selections. Input window 38 comprises a selection window 40 in which the user selects the terms to define the context group and a selection pad 42, for selecting Boolean operations to be performed on the terms.

Preferably, selection window 40 lists all the terms which appear in at least one of the documents of the corpus, and the user selects the terms form the list that will define the context group. Alternatively or additionally, the terms in selection window 40 are determined automatically, as described, for example, in an article by Feldman, et al., entitled "Text Mining at the Term Level," in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining* (1998), which is incorporated herein by reference. In the example shown in Fig. 3, the context group is defined by terms associated with "merger." Alternatively or additionally, the user may define the context group using any other parameters characterizing the documents in the corpus, including the authorship, origin, length, and date of the documents.

Preferably, an additional selection window 44 enables the user to define types of terms that will be used in generating the context graphs. Ideally it would be desirable not to limit the terms appearing in the graphs. However, in most cases, such an unlimited approach would lead to an excess of meaningless data, for example, appearances of connection words ("and," "the," etc.), in the results. Therefore, window 44 allows the user to select the terms to appear in the results. Preferably, the terms appearing in the results are chosen according to predefined groups, such as companies. personal names, etc. Alternatively or additionally, the terms allowed to appear in the results may be chosen by excluding non-interesting terms.

The user preferably chooses in a window 46 a granularity for the time axis of the documents. The granularity defines the period of time from which all documents are considered to belong to a single group. The granularity may be on the order of months, as shown in Fig. 3, or on the order of hours, days, weeks or years, or substantially any time order.

After making appropriate selections, the user preferably actuates a compute button 48, which initiates the text mining. Computer 20 searches the records which represent the documents in the context group, in order to find documents in which pairs of two different terms from among the result terms of window 44 both appear. For each pair of terms, computer 20 counts separately for each period of time the number of documents in which the pair appears. Alternatively or additionally, computer 20 assigns each pair of terms an occurrence frequency value which is based on the number of documents in which the pair appears, the number of time each of the terms appears in the documents, and/or a weight given to each term according to its importance. The search is preferably performed as described in the above-mentioned article of Feldman, Klosgen and Zilberstien. Preferably, the results are shown in a table 50, in which two columns 54 show the pairs of terms, and the rest of the columns show the number of documents for each time period.

Preferably, the rows in table 50 are sorted such that pairs which include common terms appear next to each other. Alternatively, the rows in table 50 are sorted according to the total number of documents in which the respective pairs of terms appear. Further alternatively, the rows in table 50 are sorted according to the appearance of the term pairs in a selected period, i.e., in a column or group of columns in the table. Preferably, only pairs of terms with a relatively high number of appearances are displayed in table 50, and only a predetermined number of pairs of terms are displayed. Alternatively, all pairs which

have a number of occurrences above a predefined threshold are displayed. Preferably, a button 52 allows the user to see the results in table 50 in a graphic format, as described hereinbelow.

Fig. 4A is a schematic view of a trend graph 60, in accordance with a preferred embodiment of the present invention. Graph 60 preferably represents a single column of table 50, i.e., a single period. Rows of table 50 in which the entry of graph 60 in the single column has a value above a predetermined threshold are referred to herein as active rows. Each term which appears in one the first two columns 54 of an active row is shown by a node 62 in graph 60. Each of the active rows appears as an edge 64 in graph 60. Alternatively or additionally, other rows of table 50 in which the entry at the column of graph 60 is non-zero are also considered active rows, to be represented by an edge, provided they had a value above the threshold in a previous column of table 50, typically corresponding to a preceding period. Preferably, each edge 64 is displayed along with a weight 66 which is equal to the number of documents in which the pair of terms connected by the edge appears.

Further preferably, a symbol 68 is displayed next to weight 66 designating the change in the value of the weight relative to the previous column. Alternatively, the symbol designates the change relative to an average of a number of preceding columns. For example, symbol 68 is a "<" if the weight of its edge decreased, a ">" if the weight increased, and a "*" if the weight remains substantially stable. Preferably, weights are considered to increase or decrease only if the change is larger than a predetermined factor, for example, 25%. Edges which change by a factor smaller than the predetermined factor are considered stable. Preferably, new edges and/or edges with increased weights are designated by wider lines than edges which have decreased weights. Alternatively or additionally, other sets of symbols may be used to indicate the changes in the graphs.

Fig. 4B is a schematic view of a trend graph 80 representing a period following the period represented by graph 60, in accordance with a preferred embodiment of the present invention. Preferably, nodes 62 which appear in both graph 60 and 80 are positioned in the same locations in both graphs. Therefore, space is allocated for the nodes that will appear in the graphs representing all the columns of table 50, before displaying any of the graphs. For example, empty space 70 is left in graph 60, to leave room for nodes 72 in graph 80. Thus, it is easy to follow the similarities and changes in the graphs as they are displayed,

for example, when successive graphs are displayed in sequence or in pseudo-3D geometrical superposition.

Alternatively, the positions of nodes 62 are chosen separately for each graph, arbitrarily or according to the weights of the edges 64 incident on the nodes. For example, nodes 62 having relatively higher sums of weights of the incident edges may be positioned in the center or at the top of the graph. Further alternatively or additionally, the lengths of edges 64 may be used to indicate a desired parameter. For example, the length of the edge may indicate the weight of the edge, while the thickness of the edge indicates its weight relative to one or more previous periods.

Preferably, the user can request more information by selecting areas of the graph. For example, when the user double-clicks on one of edges 64, a window may open with a bar graph, a table or any other indication which shows the weights of the edge as a function of time. Alternatively or additionally, the documents contributing to the selected edge may be listed, allowing the user to read the documents and judge their relevance. Further preferably, the user may request to see the graphs as they change over time in an animation sequence.

Fig. 5 is a schematic view of a comparison graph 100, in accordance with a preferred embodiment of the present invention. Graph 100 compares text mining results of recipe documents from different document groups, for example, documents from two different countries. Each major ingredient in the recipes is designated by a node 102. Nodes 102 which appear together in more than a predetermined threshold number of documents are connected by an edge 104. Each edge is marked with two values, corresponding to appearance of the associated terms in documents from the two different countries. Preferably, the values indicate the percentage of documents from the respective country in which the pair of ingredients connected by the corresponding edge 104 both appear. Alternatively or additionally, the edges 104 are marked with the absolute number of documents. Preferably, edges 106 which correspond to combinations that are more popular in country #1 are displayed differently from edges 108 for combinations which are more popular in country #2.

Alternatively or additionally, the edges and values for each country may be displayed in different colors. Thus, it is possible to compare documents from more than two groups. Further alternatively or additionally, only a single value designating the

difference between the values of different document groups is displayed with each edge. Preferably, the user may select which type of display is desired.

Fig. 6 is a schematic view of a graphic interface 120, showing sample graphs 122, 124, 126, and 128, for displaying results generated in part using some of the techniques described hereinabove, in accordance with a preferred embodiment of the present invention. Graphs 122 and 124 are, respectively, a "single-term"-centered graph and a bar graph, in which the relationship between a single term ("Microsoft") and a set of other terms ("IBM," Sun," etc.) is quantitatively displayed. The quantitative relationship shown in graphs 122 and 124 may comprise, for example, the number of news articles containing both the term "Microsoft" and each of the other listed terms during a specified time period. Using the same analysis as that which generated graphs 122 and 124, graph 126 is displayed to show the most significant relationships among all of the displayed terms. By contrast, graph 128 shows the number of appearances of the term "Microsoft," irrespective of the other companies, during a five week period extending from April 10 to May 15.

Preferably, a slide-bar 130 is provided with interface 120, which enables the user to move an enhanced slide-piece 132 between two points on an axis of interest, e.g., time. Slide-bars which perform this limited function are widely available, for example, in Microsoft Windows 98. In prior art slide-bars, the slide-piece is typically moved to indicate, for example, a location in a document, a time, or a color from a range of pages, times, or colors, respectively.

In this embodiment, the length of enhanced slide-piece 132, i.e., the distance between points 134 and 136 in Fig. 6, provides the user with additional information about a parameter of interest. For example, slide-bar 130 in the embodiment shown in Fig. 6 represents a set of relevant news articles spanning one year. The length of enhanced slide-piece 132, as shown, is five weeks, i.e., approximately one tenth the total length of slide-bar 130. Preferably, as the user moves the enhanced slide-piece along the slide-bar, graphs 122, 124, and 126 are continually updated responsive to whatever news articles are contained in a five-week period which is "covered" by the slide-piece.

Further preferably, and completely unlike any slide-bar known in the art, the user is enabled to modify the length of the enhanced slide-piece in real time, so as to cause computer 20 to change the set of articles used in generating the graphs accordingly. For example, dashed lines 148 show a former setting of the slide-piece, in which

approximately twelve weeks were represented by the slide-piece. Preferably, the user uses a mouse to grab onto the left side 144 or right side 146 of enhanced slide-piece 132, and changes its length, typically in a manner analogous to the way objects are re-sized in a Windows environment. Notably, however, neither Windows nor any other software provides the improved and intuitive position control provided by enhanced slide-piece 132.

In light of this description of the operation of slide-piece 132, many applications not related to a time axis will become obvious to one skilled in the art. For example, scrolling through a document by moving slide-piece 132 could be enhanced by a "zoom" feature, effectively enabled by changing the size of the slide-piece. Alternatively, whereas a slide-bar which uses prior art technology would allow the user to select a single color from a spectrum, a user of embodiments of the present invention would be additionally enabled to select a range of neighboring colors in an intuitive fashion.

It will be understood by one skilled in the art that aspects of the present invention described hereinabove can be embodied in a computer running software, and that the software can be stored in tangible media, e.g., hard disks, floppy disks or compact disks, or in intangible media, e.g., in an electronic memory, or on a network such as the Internet.

It will be appreciated by persons skilled in the art that the present invention is not limited to what has been particularly shown and described hereinabove. Rather, the scope of the present invention includes both combinations and subcombinations of the various features described hereinabove, as well as variations and modifications thereof that are not in the prior art, which would occur to persons skilled in the art upon reading the foregoing description.

CLAIMS

1. A method for visualizing variations in a corpus of information, including a plurality of information entries which are divided into a plurality of sub-groups according to a differentiating parameter of the entries, comprising:

for each of the entries, extracting characteristics of information contained therein;

finding pairs of different characteristics that appear together in at least one of the entries;

determining an occurrence value for each of the pairs of characteristics in each sub-group in which both of the characteristics appear;

comparing the occurrence values of at least some of the pairs of characteristics for at least two of the sub-groups; and

providing an indication of the comparative occurrence values of the pairs.

2. A method according to claim 1, wherein the entries comprise text documents, and wherein the characteristics comprise terms appearing in the documents.

3. A method according to claim 1, wherein determining the occurrence value comprises counting the number of entries in which the pair appears.

4. A method according to claim 1, wherein finding the pairs of characteristics comprises finding pairs of characteristics which appear together in at least a predetermined number of the entries.

5. A method according to claim 1, wherein finding the pairs of characteristics comprises finding pairs of characteristics which appear together in at least two of the sub-groups.

6. A method according to claim 1, wherein extracting the characteristics comprises automatically mining the corpus to extract characteristics therefrom.

7. A method according to any one of claims 1-6, wherein the differentiating parameter defines an order, and wherein comparing the occurrence values comprises comparing the occurrence values in a first sub-group with the occurrence values in one or more previous sub-groups in the order.

8. A method according to claim 7, wherein comparing the occurrence values comprises comparing the occurrence values in the first sub-group with the occurrence values in a closest previous sub-group.

9. A method according to claim 7, wherein comparing the occurrence values comprises comparing the occurrence values in the first sub-group with an average of the occurrence values in the one or more previous sub-groups.

10. A method according to claim 7, wherein providing the indication comprises displaying a symbol which indicates a measure of evolution in the occurrence value in the first sub-group relative to the occurrence values in the one or more previous sub-groups in the order.

11. A method according to any one of claims 1-6, wherein providing the indication comprises displaying a table.

12. A method according to any one of claims 1-6, wherein providing the indication comprises displaying a graph.

13. A method according to claim 12, wherein displaying the graph comprises displaying a graph in which each term is represented by a node, the pairs of characteristics that are found are represented by edges, and substantially each edge is associated with the indication of the comparative appearance of the respective pair.

14. A method according to claim 13, wherein displaying the graph comprises displaying with substantially each edge a weight of the edge, which equals the occurrence value of the respective pair in a first sub-group.

15. A method according to claim 13, wherein displaying the graph comprises displaying the graph such that the lengths of the edges represent the occurrence value of the respective pair in a first sub-group.

16. A method according to claim 12, wherein displaying the graph comprises displaying for each two sub-groups a graph which compares the occurrence values in the two sub-groups.

17. A method according to claim 16, wherein displaying the graph for each two sub-groups comprises displaying the graphs such that nodes which represent the same term are displayed in substantially the same relative location.

18. A method according to claim 16, wherein the graphs of each two sub-groups are displayed as an animation sequence.

19. A method according to claim 12, wherein displaying the graph comprises displaying a plurality of superimposed graphs, each of which represents the appearances of the pairs in a different sub-group.

20. A method according to claim 19, wherein displaying the plurality of superimposed graphs comprises displaying each of the graphs in a different color.

21. A method according to any one of claims 1-6, wherein providing the indication of the comparative values of the pairs comprises providing an indication wherein pairs having a characteristic in common are grouped together.

22. Apparatus for visualizing variations in a corpus of information including a plurality of information entries which are divided into a plurality of sub-groups according to a differentiating parameter of the entries, comprising:

a processor which finds pairs of characteristics which appear together in at least one of the documents, determines an occurrence value for each of the pairs of characteristics in each sub-group in which both of the characteristics appear, and compares the occurrence values of at least some of the pairs of characteristics for at least two of the sub-groups; and

a display which displays an indication of the comparative occurrence values of the pairs.

23. Apparatus according to claim 22, wherein the entries comprise text documents, and wherein the characteristics comprise terms appearing in the documents.

24. Apparatus according to claim 22, wherein the occurrence value comprises the number of entries in which the pair appears.

25. Apparatus according to claim 22, wherein the processor finds those pairs of characteristics which appear together in at least a predetermined number of entries.

26. Apparatus according to claim 22, wherein the processor finds those pairs of characteristics which appear together in at least two of the sub-groups.

27. Apparatus according to claim 22, wherein the processor finds characteristics selected from a group of automatically determined characteristics.

28. Apparatus according to any one of claims 22-27, wherein the differentiating parameter defines an order and wherein the processor compares the occurrence values in a

first sub-group with the occurrence values in one or more previous sub-groups in the order.

29. Apparatus according to claim 28, wherein the processor compares the occurrence values in the first sub-group with the occurrence values in a closest previous sub-group.

30. Apparatus according to claim 28, wherein the processor compares the occurrence values in the first sub-group with an average of the occurrence values in the one or more previous sub-groups.

31. Apparatus according to claim 28, wherein the display displays a symbol which indicates a measure of evolution in the occurrence values in the first sub-group relative to the occurrence values in the one or more previous sub-groups in the order.

32. Apparatus according to any one of claims 22-27, wherein the display displays a table.

33. Apparatus according to any one of claims 22-27, wherein the display displays a graph.

34. Apparatus according to claim 33, wherein each node in the graph represents a term and each edge represents a found pair of characteristics, and substantially each edge is associated with the indication of the comparative appearance of the respective pair.

35. Apparatus according to claim 34, wherein the graph comprises with substantially each edge a weight of the edge which equals the occurrence value of the respective pair in a first sub-group.

36. Apparatus according to claim 34, wherein the graph comprises a graph in which the lengths of the edges represent the occurrence values of the respective pairs in a first sub-group.

37. Apparatus according to claim 33, wherein the graph comprises a plurality of graphs each of which compares the occurrence values of the pairs in two sub-groups.

38. Apparatus according to claim 37, wherein the plurality of graphs comprise graphs such that nodes which represent the same term are displayed in substantially the same relative location.

39. Apparatus according to claim 37, wherein the plurality of graphs are displayed as an animation sequence.

40. Apparatus according to claim 33, wherein the graph comprises a plurality of superimposed graphs each of which represents the occurrence values of the pairs in a different sub-group.

41. Apparatus according to claim 40, wherein the plurality of superimposed graphs comprise a plurality of superimposed graphs in which each of the graphs is displayed in a different color.

42. Apparatus according to any one of claims 22-27, wherein the display displays the pairs such that pairs which have common characteristics are grouped together.

43. A method for selecting a range of values of a variable, comprising:

providing a graphic user interface on a display, including a slide-piece that has an initial dimension and is translatable along an axis representing the variable such that each position of the slide-piece along the axis corresponds to a given value of the variable;

positioning the slide-piece at a first position on the axis, so as to indicate a first value of the variable; and

changing the dimension of the slide-piece so as to indicate a second value of the variable, whereby the first and second values of the variable define the selected range.

44. A method according to claim 43, wherein changing the dimension of the slide-piece comprises changing a length of the slide-piece along the axis.

45. A method according to claim 43 or claim 44, wherein the first and second values of the variable comprise the extrema of the range.

46. A computer program product for visualizing variations in a corpus of information, including a plurality of information entries which are divided into a plurality of sub-groups according to a differentiating parameter of the entries, the documents including text, the program having computer-readable program instructions embodied therein, which instructions, when read by a computer, cause the computer to:

for each of the entries, extract characteristics of information contained therein;

find pairs of different characteristics that appear together in at least one of the entries;

determine an occurrence value for each of the pairs of characteristics in each sub-group in which both of the characteristics appear;

compare the occurrence values of at least some of the pairs of characteristics for at least two of the sub-groups; and

provide an indication of the comparative occurrence values of the pairs.

47. A computer program product for selecting a range of values of a variable, the program having computer-readable program instructions embodied therein, which instructions, when read by a computer, cause the computer to:

provide a graphic user interface on a display, including a slide-piece that has an initial dimension and is translatable along an axis representing the variable such that each position of the slide-piece along the axis corresponds to a given value of the variable;

position the slide-piece at a first position on the axis, so as to indicate a first value of the variable; and

change the dimension of the slide-piece so as to indicate a second value of the variable, whereby the first and second values of the variable define the selected range.

48. A method for extracting data from a corpus of information including a plurality of information entries, each entry being assigned to one or more sub-groups according to a differentiating parameter of the entries, comprising:

for a first one of the entries in a first one of the sub-groups, extracting a characteristic of information contained therein;

for a second one of the entries in a second one of the sub-groups, extracting the same characteristic of information;

automatically determining respective first and second occurrence values corresponding to the characteristic in the first and second sub-groups; and

providing an indication of the occurrence values.

49. A method according to claim 48, wherein providing the indication comprises providing a visual indication of the occurrence values.

50. A method according to claim 48, wherein the entries comprise text documents, and wherein the characteristic comprises a term appearing in the documents.

51. A method according to any one of claims 48-50, wherein the differentiating parameter comprises a sequence.

52. A method according to claim 51, wherein the sequence comprises a time sequence.

53. Apparatus for extracting data from a corpus of information including a plurality of information entries, each entry being assigned to one or more sub-groups according to a differentiating parameter of the entries, comprising:

a processor, which (a) for a first one of the entries in a first one of the sub-groups, extracts a characteristic of information contained therein, (b) for a second one of the entries in a second one of the sub-groups, extracts the same characteristic of information, and (c) automatically determines respective first and second occurrence values corresponding to the characteristic in the first and second sub-groups; and

a display, which provides an indication of the occurrence values.

54. Apparatus according to claim 53, wherein the display provides a visual indication of the occurrence values.

55. Apparatus according to claim 53, wherein the entries comprise text documents, and wherein the characteristic comprises a term appearing in the documents.

56. Apparatus according to any one of claims 53-55, wherein the differentiating parameter comprises a sequence.

57. Apparatus according to claim 56, wherein the sequence comprises a time sequence.

58. A computer program product for extracting data from a corpus of information, including a plurality of information entries, each entry being assigned to one or more sub-groups according to a differentiating parameter of the entries, the program having computer-readable program instructions embodied therein, which instructions, when read by a computer, cause the computer to:

for a first one of the entries in a first one of the sub-groups, extract a characteristic of information contained therein;

for a second one of the entries in a second one of the sub-groups, extract the same characteristic of information;

automatically determine respective first and second occurrence values corresponding to the characteristic in the first and second sub-groups; and

provide an indication of the occurrence values.

59. A product according to claim 58, wherein providing the indication comprises providing a visual indication of the occurrence values.

60. A product according to claim 58, wherein the entries comprise text documents, and wherein the characteristic comprises a term appearing in the documents.

61. A product according to any one of claims 58-60, wherein the differentiating parameter comprises a sequence.

62. A product according to claim 61, wherein the sequence comprises a time sequence.

FIG. 1

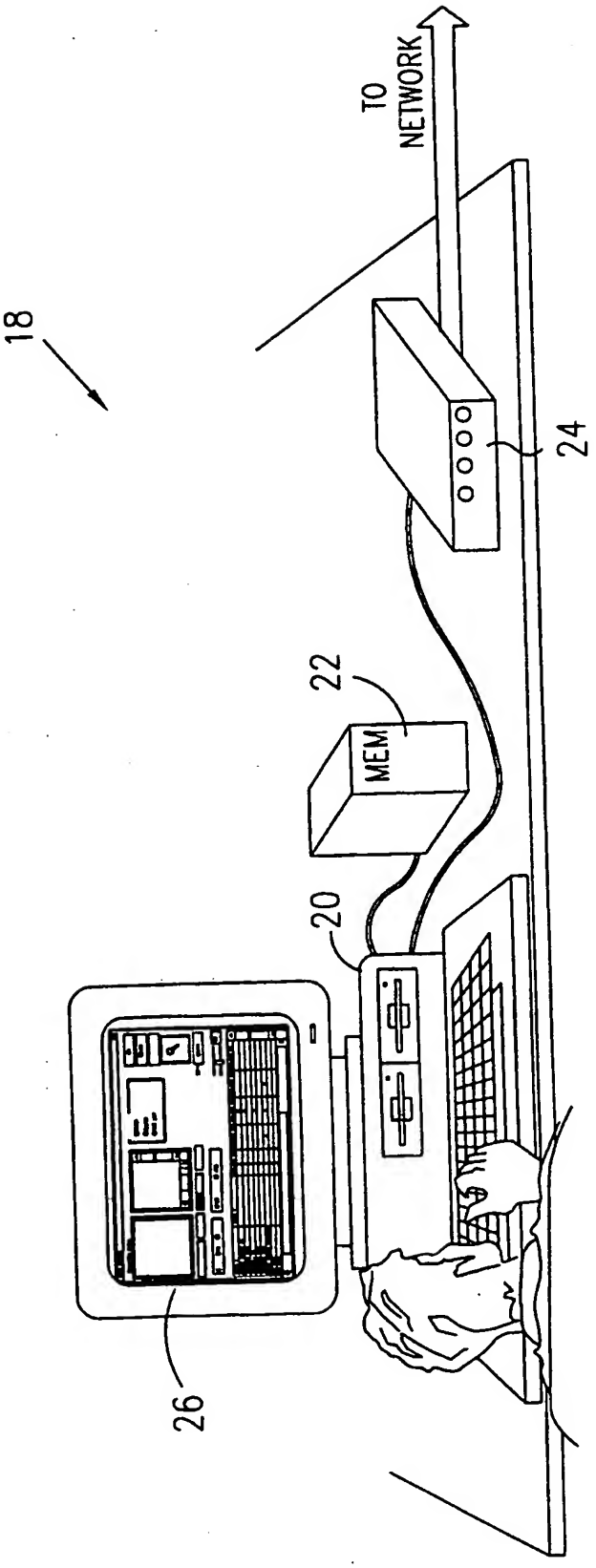


FIG. 2

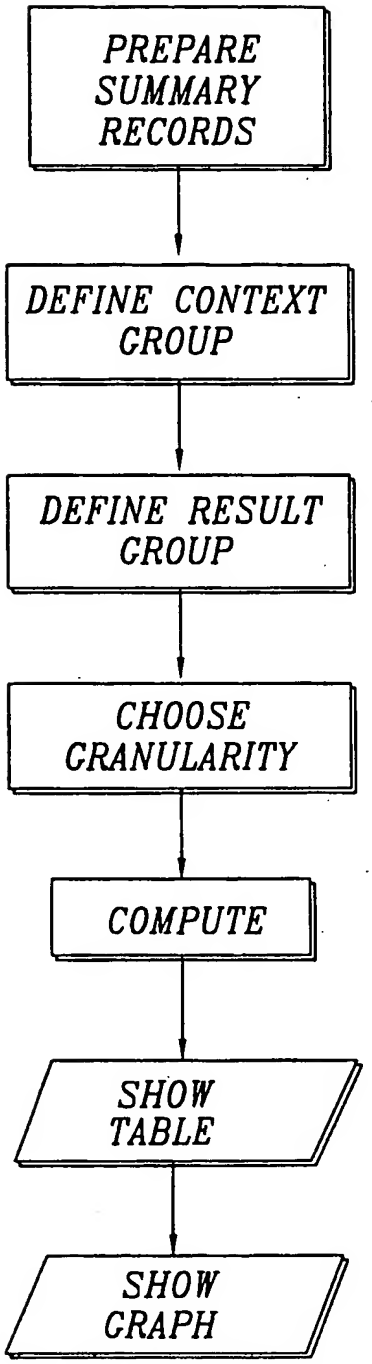

PREPARE SUMMARY RECORDS
DEFINE CONTEXT GROUP
DEFINE RESULT GROUP
CHOOSE GRANULARITY
COMPUTE
SHOW TABLE
SHOW GRAPH

FIG. 3

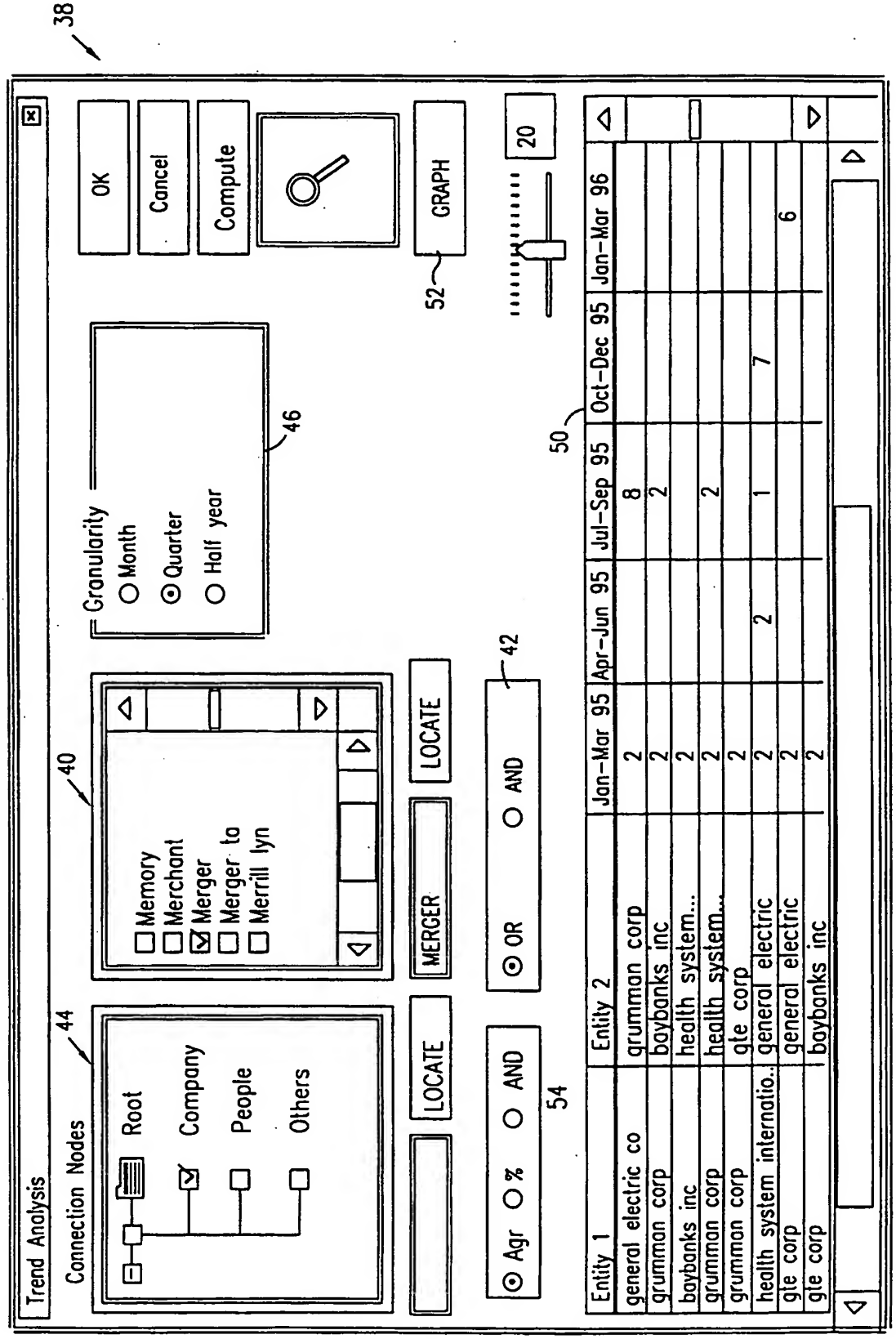

38
Trend Analysis
Connection Nodes
44
Root
Company
People
Others
40
Memory
Merchant
Merger
Merger to
Merrill lyn
LOCATE
LOCATE
MERGER
Agr
%
AND
OR
AND
42
54
Granularity
Month
Quarter
Half year
46
OK
Cancel
Compute
GRAPH
52
20
50
Entity 1
Entity 2
Jan-Mar 95
Apr-Jun 95
Jul-Sep 95
Oct-Dec 95
Jan-Mar 96
general electric co
grumman corp
2
8
grumman corp
baybanks inc
2
2
baybanks inc
health system...
2
grumman corp
health system...
2
2
grumman corp
gte corp
2
health system internatio..
general electric
2
2
1
7
gte corp
general electric
2
6
gte corp
baybanks inc
2

FIG. 4A

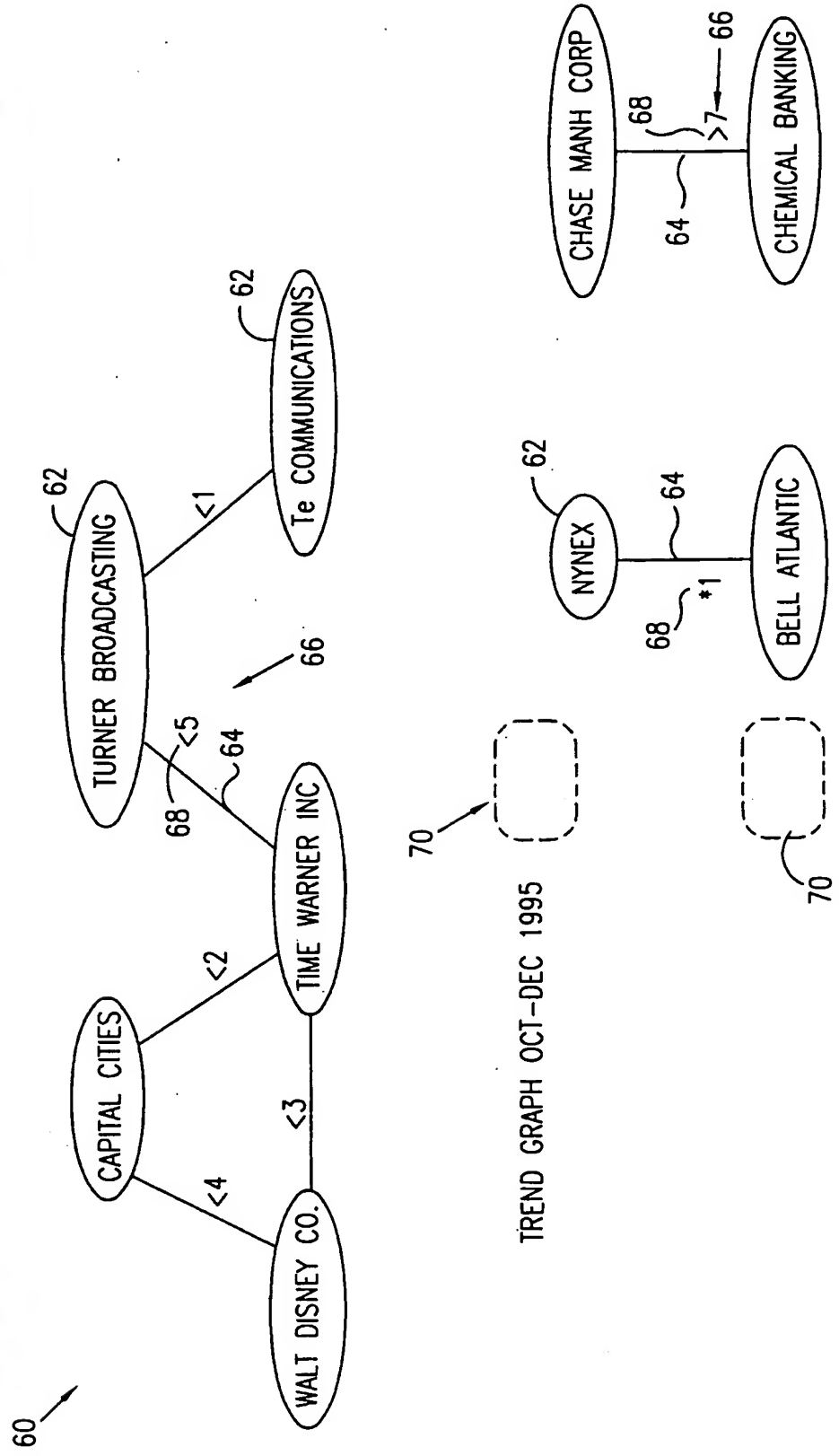

60
TURNER BROADCASTING
62
Te COMMUNICATIONS
62
<1
<5
68
64
66
TIME WARNER INC
CAPITAL CITIES
<2
<4
<3
WALT DISNEY CO.
70
TREND GRAPH OCT-DEC 1995
CHASE MANH CORP
68
>7
66
64
CHEMICAL BANKING
NYNEX
62
64
68
*1
BELL ATLANTIC
70

FIG. 4B

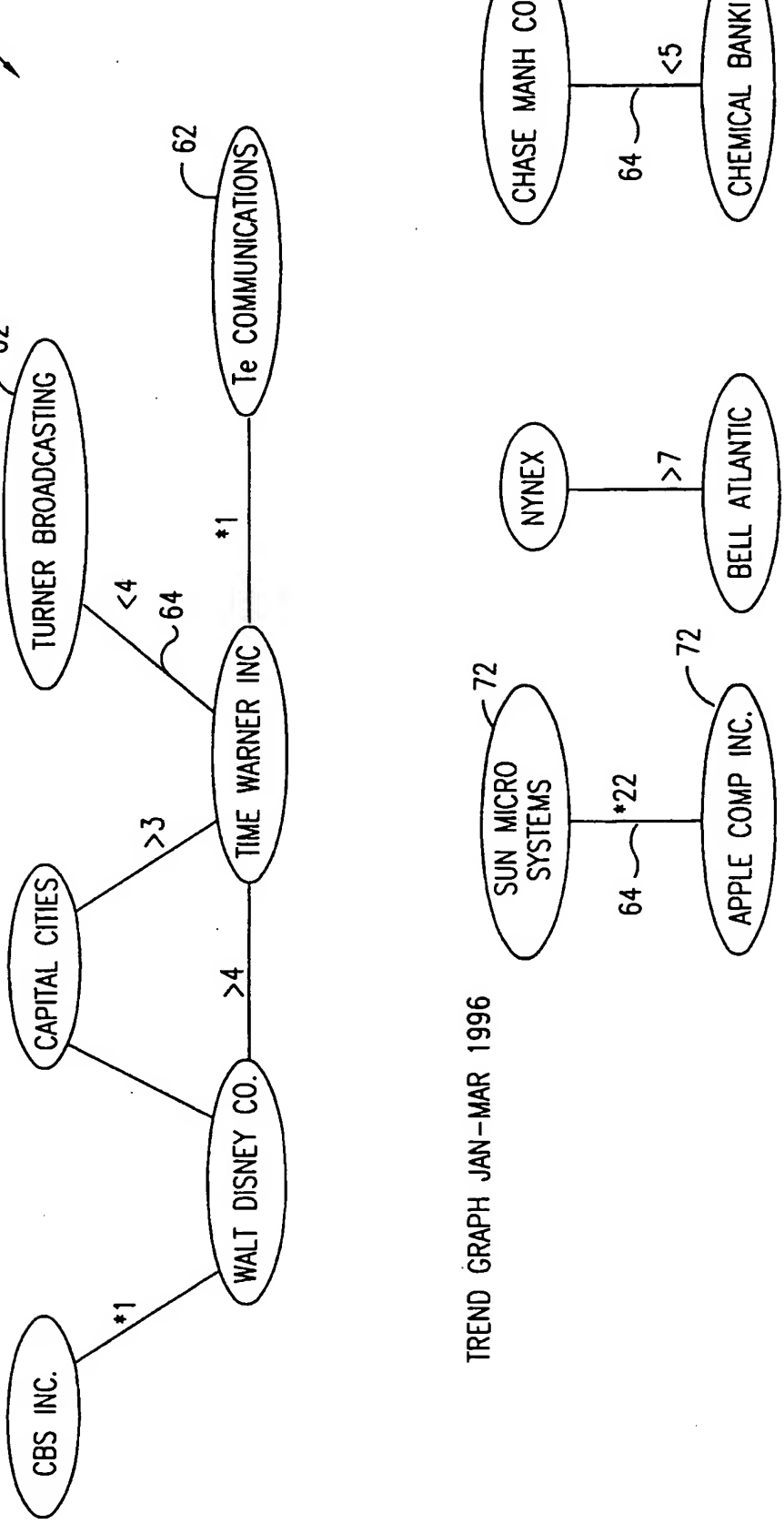

80
CBS INC.
*1
WALT DISNEY CO.
>4
CAPITAL CITIES
>3
TIME WARNER INC
<4
64
TURNER BROADCASTING
62
*1
Te COMMUNICATIONS
62
SUN MICRO SYSTEMS
72
64
*22
APPLE COMP INC.
72
NYNEX
>7
BELL ATLANTIC
CHASE MANH CORP
64
<5
CHEMICAL BANKING
TREND GRAPH JAN-MAR 1996

FIG. 5

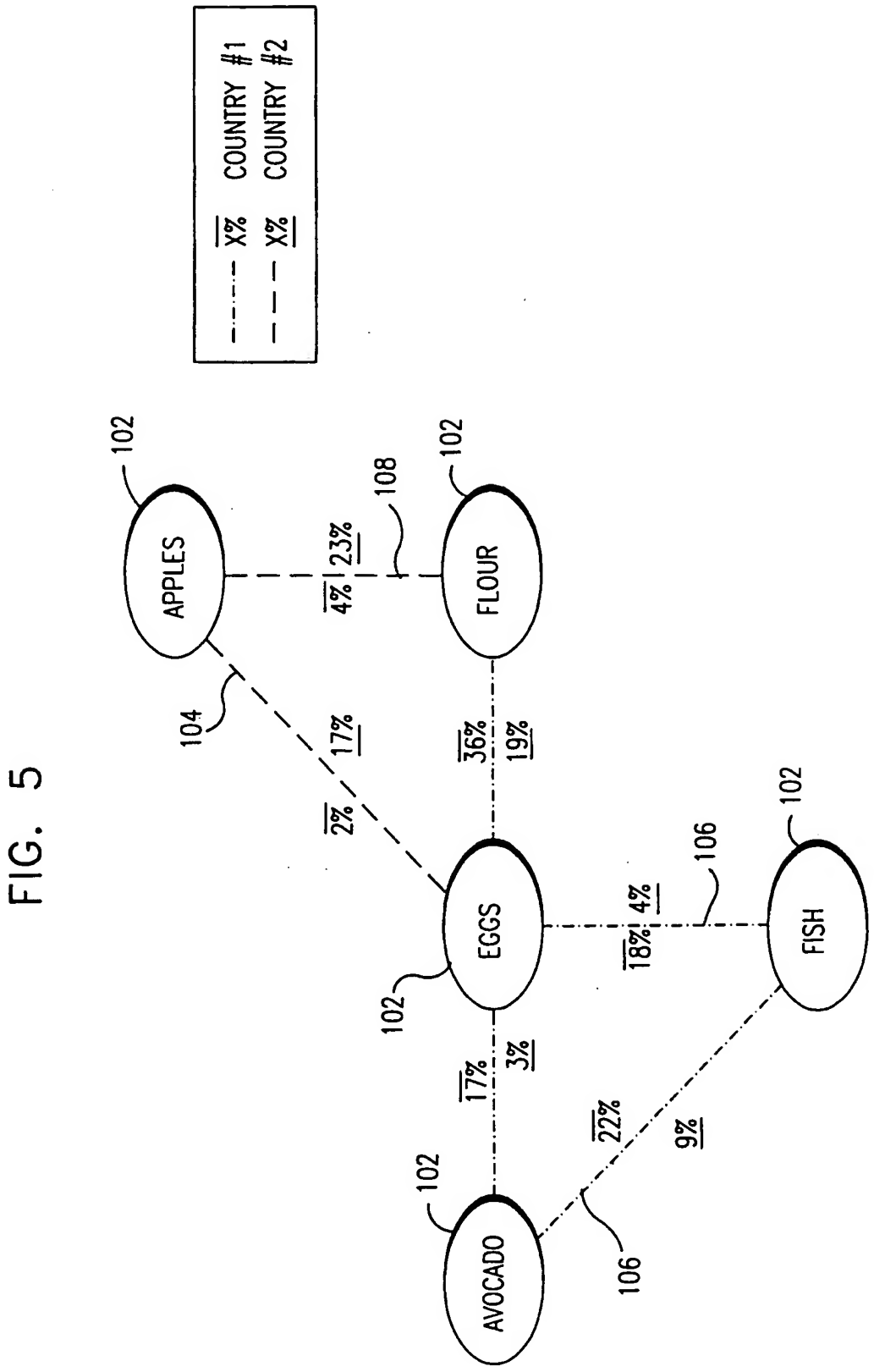
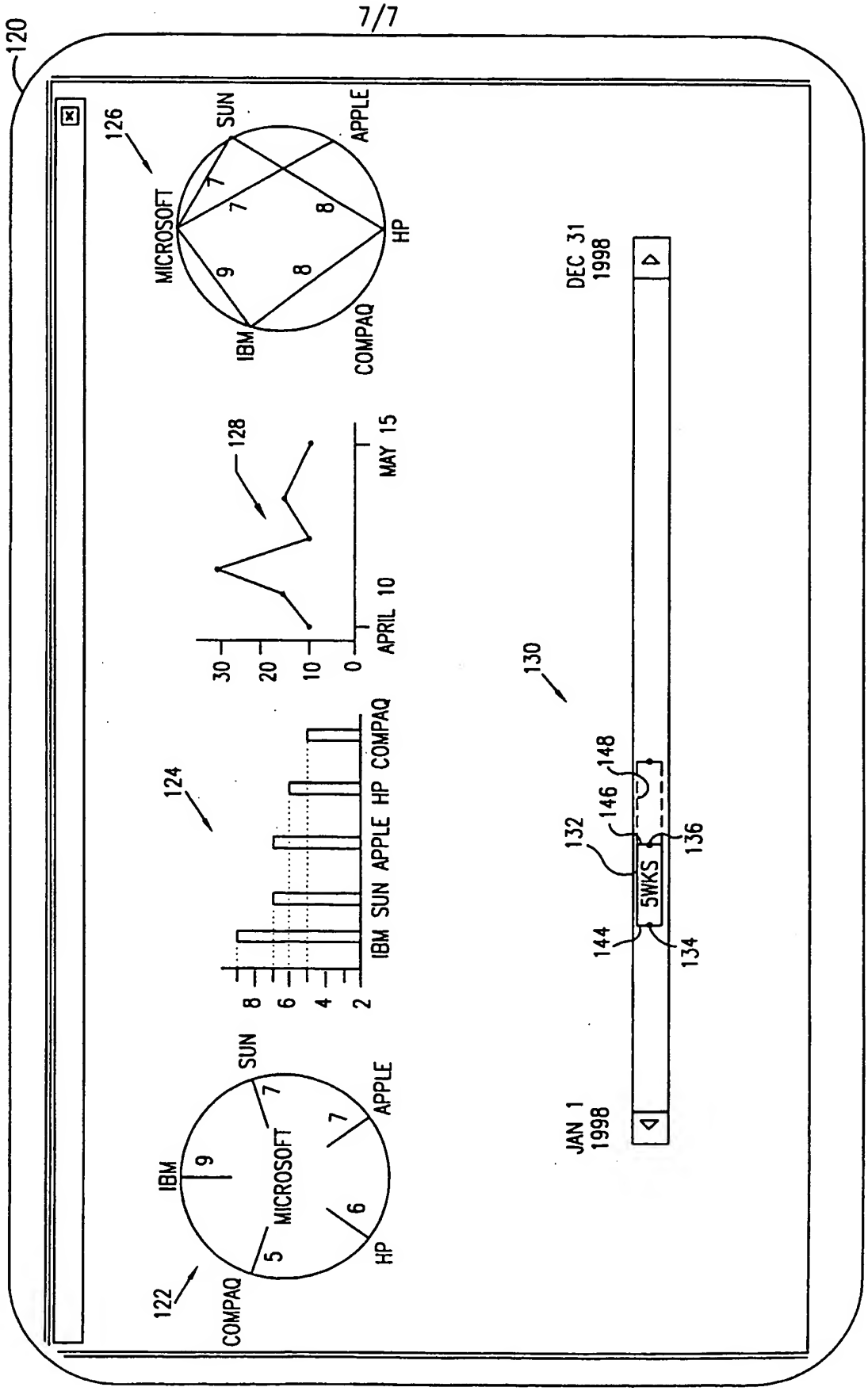

X% COUNTRY #1
X% COUNTRY #2
APPLES
FLOUR
EGGS
FISH
AVOCADO
102
104
106
108
4% 23%
2% 17%
36% 19%
18% 4%
17% 3%
22% 9%

FIG. 6


120
126
MICROSOFT
SUN
APPLE
HP
IBM
COMPAQ
7
7
8
9
8
128
30
20
10
0
APRIL 10
MAY 15
124
IBM SUN APPLE HP COMPAQ
8
6
4
2
122
COMPAQ
IBM
SUN
APPLE
HP
MICROSOFT
9
7
7
6
5
130
DEC 31
1998
JAN 1
1998
132
146
148
136
144
134
5WKS